Genomics 100 (2012) 282–288

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

RCM: A novel association approach to search for coronary artery disease genetic related metabolites based on SNPs and metabolic network

Xu Li [a,1], Lina Chen [a,*,1], Liangcai Zhang [a,1], Wan Li [a], Xu Jia [a], Weiguo Li [a], Xiaoli Qu [a], Jingxie Tai [a], Chenchen Feng [a], Fan Zhang [a], Weiming He [b]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Hei Longjiang Province, China
[b] Institute of Opto-electronics, Harbin Institute of Technology, Harbin, Hei Longjiang Province, China

ARTICLE INFO

Article history:
Received 19 March 2012
Accepted 20 July 2012
Available online 29 July 2012

Keywords:
Coronary artery disease
Single nucleotide polymorphisms
Genetic risks
Metabolic network
Organelles

ABSTRACT

Integration of genetic and metabolic network holds promise for providing insight into human disease. Coronary artery disease (CAD) is strongly heritable, but the heritability of metabolic compounds has not been evaluated in human metabolic context. Here we performed a genetic-based computational approach within eight sub-cellular networks from Edinburgh Human Metabolic Network to identify significant genetic risk compounds (SGRCs) of CAD. Our results provide the evidence that the high heritabilities of SGRCs played an important role in CAD pathogenesis. Besides, SGRCs were discovered to be strongly associated with lipid metabolism. We also established a possible disease-causing reference table to decipher genetic associations of SGRCs with CAD. Comparing with traditional method, RCM experienced better performance in CAD genetic risk compounds' identification. These findings provided novel insights into CAD pathogenesis from a genetic perspective.

1. Introduction

Coronary artery disease (CAD) is rapidly becoming the leading cause of death in the world, and is considered to be closely relevant with genetics and biological metabolism [1]. Current studies on CAD are basically relying on detecting the variation of CAD related metabolic bio-markers and identifying their functions for clinical diagnosis and treatment [2]. Understanding the role of metabolic compounds and their interaction with environmental factors in CAD is the key to the development of safe and efficient therapies, to diagnosis and to prevention [3].

The genetic predilection of CAD is well established. For example, family history has been proved to be one of the independent risk factors for CAD [4]. Many metabolic risk factors for CAD need further explorations. Genome-wide association studies (GWAS) have great power in CAD genetic risk researches [5–8]. Chen and Zhang [9] utilized human metabolic pathways and high-throughput SNP datasets to find genetic risk factors which are related with complex diseases. Suhre and Shin [3] used a GWAS with non-targeted metabolites to identify genetic loci associated metabolites. In both studies, the majority of the analytes related with CAD genetics remained unidentified. Besides, researches on CAD that combine GWAS with metabolites are still limited. All published GWA studies focused mainly on the discovery of novel genes, while researchers paid little attention to the identification of disease-related metabolites in part for lack of robust frameworks (e.g., metabolic network) that consist of detailed intracellular relationships between genes and metabolites. As one of the most authoritative human metabolic network, the Edinburgh Human Metabolic Network (EHMN) has been widely used in complex diseases [10–12]. Aleksej and Zelezniak et al. used gene expression profiles and proposed a computational method to identify Type 2 Diabetes related bio-markers in EHMN [13]. This network contains comprehensive cascade connections of enzymes, genes, reactions and metabolites, which gives us an opportunity to study CAD pathogenesis from the viewpoints of genetic factors and metabolic network context.

In this paper, we proposed a genetic-based computational approach named risk compounds mining (RCM) to identify significant genetic risk compounds (SGRCs) of CAD by appropriately quantifying high-throughput SNP datasets and biologic network context and fusing them. Due to genetic differentiation between organelles, our approach was applied to the identification of organelle-specific SGRCs in eight sub-cellular networks (cytoplasm (C), extracellular space (E), mitochondria (M), Golgi apparatus (G), endoplasmic reticulum (R), lysosome (L), peroxisome (X) and nucleus (N)), respectively. After finding CAD genetic risk compounds, we adopted CAD related genes to validate the feasibility and efficiency of our approach. Besides, by literature search, many SGRCs in our study were confirmed to be traditional risk factors of CAD. Furthermore, we integrated three widely accepted

Abbreviation: CAD, Coronary artery disease; SGRCs, Significant genetic risk compounds; EHMN, Edinburgh Human Metabolic Network.

* Corresponding author at: Harbin Medical University, China. Fax: +86 451 86669617.
E-mail address: chenlina@ems.hrbmu.edu.cn (L. Chen).

[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.07.013

databases to build a comprehensive CAD pathogenic function category of our SGRCs. These results provided the evidence that the high heritabilities of SGRCs in our study play an important role in pathogenesis of CAD and the newly discovered metabolites could possibly serve as bio-markers for CAD clinical diagnosis in future researches.

2. Results

2.1. The distribution of CAD related SGRCs

In this paper, we proposed the RCM method (Materials and methods) to screen CAD genetic risk related metabolites in EHMN. However, in different organelles, the same metabolic reaction may be catalyzed by different enzymes [21], which suggests that each compound might have different genetic risk values in different organelles. The scoring method in our study was based on the fact that enzymes as biological catalysts of metabolic reactions have high specificity [22]. As a result, we utilized CAD high-throughput SNP datasets and EHMN to screen SGRCs in eight sub-networks, respectively. Totally, 209 SGRCs have been screened out (additional files Table S1); we found most SGRCs concentrated on three organelles: mitochondria, endoplasmic reticulum and peroxisome (Fig. 1A). Mitochondria, as the leading factor of coronary artery disease [23], include 39 SGRCs (19%); there are 60 SGRCs (29%) presented in the endoplasmic reticulum, which is coincident with the fact that the endoplasmic reticulum stress is the main reason for atherosclerosis [24]; 48 SGRCs (23%) were located in the peroxisome and this organelle is associated with the prevalence of CAD inflammation [25]. Besides, according to the KEGG classification standard [26], all SGRCs were participated in 74% of the human metabolic pathways (additional files Table S5) and encountered in five basic metabolic processes (Fig. 1B). Lipid metabolism contains the most SGRCs (42%), followed by amino acid metabolism (26%), Vitamin and Cotactor metabolism (13%), Carbohydrate metabolism (13%) and Nucleotide metabolism (6%). As described above, most of the SGRCs seemed to be present in different organelles, indicating that SGRCs have their specific genetic risks in different cellular organs.

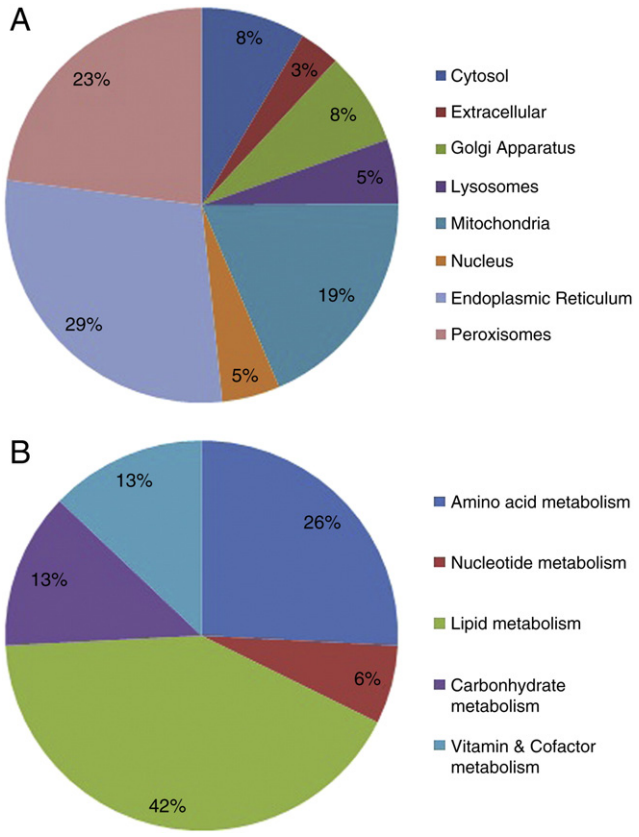

Fig. 1. The distribution of CAD SGRCs identified by RCM method. (A) The pie chart showed the distributions of SGRCs among eight sub-cellular networks. (B) The pie chart showed the distribution of SGRCs among five basic metabolic categories.

2.2. The hereditary of CAD genes and metabolites

However, besides metabolites, the heritability of CAD related genes is another important factor which can't be ignored in the research of CAD genetics [27]. Having obtained these SGRCs, we sought to further analyze the genetic risk of CAD. CAD related genes have been discovered in recent studies comprehensively [16], and we considered combining them with SGRCs in the analysis of CAD's genetic risk. First, we were interested in the issue of whether these disease genes of CAD have genetic risks through our method. We totally obtained 318 CAD related genes from the CAD gene database [16], and there are 82 CAD related genes present in EHMN (additional files Table S4). With the RCM method (Materials and methods) used here, we found the risk value of CAD related genes is generally higher than the background (5.304/3.221; Wilcoxon test; $P<0.05$). As expected, the results were consistent with previous studies that CAD related genes are heritable [28]. Interestingly, 32 metabolic compounds (additional files Table S2) regulated by CAD related genes have been found to be SGRCs, and their average risk value is generally higher than that of all the SGRCs (Wilcoxon test, $P<0.05$). Furthermore, through functional enrichment analysis, 32 SGRCs are mainly concentrated in lipid and lipoprotein metabolism, glucose metabolism, homocysteine metabolism, integrity of endothelial cells, abnormal vascular smooth muscle cells, and oxidation–reduction (additional files Table S2). The results above showed that part of CAD related genes are closely correlated with the risk factors in CAD, and the SGRCs which are directly regulated by those genes can possibly serve as hereditary factors in the progress of CAD.

2.3. Literature retrieval and functional classification of SGRCs

We adopted literature retrieval to look for the evidence of the correlation between SGRCs and CAD pathogenesis. To all 209 SGRCs, 103 SGRCs (49.3%) had been reported in the literature (additional files Table S7). For instance, Acylglycerol (risk value (RV) = 269.4390, P = 7.24E−07, C), Diacylglycerol (RV = 269.43900, P = 7.24E−07, C) and Glycerol lipid (RV = 68.5722, P = 2.55E−11, C) are the most commonly used indicators in CAD clinical detection [29]. Elevated levels of Triglyceride (RV = 269.4390, P = 7.24E−07, C) in blood is often accompanied by coagulation dysfunction and abnormal lipid metabolism, which are important factors in vascular damages [30]. An increase of the NADH + (RV = 4.2915, P = 1.68E−07, M)/NADPH (RV = 4.2915, P = 2.05E−09, M) level would accelerate oxygen metabolism, leading to endothelial dysfunction, inflammation and extracellular matrix deposition, and eventually accelerate the vascular atherosclerosis progress [31–33]. High levels of lipoprotein which transports cholesterol in the blood are thought to be associated with increased risk of CAD and atherosclerosis [1], while the low-density lipoprotein can effectively reduce the incidence of CAD [34]. Diacylglycerol (DAG) (RV = 269.4390, P = 7.24E−07, C) could potentially affect lipid sensitivity via activation of serine/threonline kinases or alterations in phospholipid membrane composition, both of which could lead to lipid synthesis [35]. All together, identification of these SGRCs known to be involved in coronary artery disease pathogenesis provided further evidence for the validation of our approach.

Nevertheless, there have been no comprehensive classifications for function categories of metabolites and their roles in CAD. Therefore, we considered to build a resource of our SGRCs mapping with CAD related functional categories for better understanding the role of genetic risk factors in the CAD pathogenesis. Here, we used three widely accepted databases as data sources of functional categories, including Pubchem [36], MBRole [20] and KEGG [26]. By integrating

the classification criterion of these databases, we divided the pathogenic mechanism of CAD into 12 categories (Table 1). Furthermore, we also classified literature confirmed SGRCs into 12 categories. Interestingly, nearly half of SGRCs (45.6%) were classified into the categories of lipid and lipoprotein metabolism and oxidation–reduction (Table 1). Current studies indicated the leading cause of cardiovascular disease is lipid metabolism disorder [37,38]. Homocysteine (RV = 15.4722, P = 0.0015, N) and Trichloroethene (RV = 45.6722, P = 0.0201, R) are related to endothelial dysfunction [39,40], therefore we classified them into endothelial integrity category. Moreover, some plasminogen activator inhibitor like GDP (RV = 14.3203, P = 3.24E−13, G) and GDP-L-fucose (RV = 14.7214, P = 4.75E−08, G) are essential to thrombosis [41]. As is shown in Table 2, the rest of SGRCs are specifically distributed in other CAD functional categories, such as glucose metabolism, renin–angiotensin system, gender difference, vascular smooth muscle cell abnormalities, homocysteine metabolism, immune and inflammation and Purine alkaloids. The above results have clearly classified part of the SGRCs into 12 CAD functional categories, which suggested that some of our SGRCs have essential roles in the mechanisms of the CAD pathogenesis. In this work, there were also some SGRCs which have not been confirmed by current studies. Interestingly, many of these non-literature confirmed SGRCs belong to lipid metabolism and glucose metabolism. A prime example is that Isopentenyl diphosphate (RV = 9.2088, P = 0.0472, X), 3-Oxododecanoyl-CoA (RV = 8.0135, P = 0.0381, X) and (S)-Hydroxydecanoyl-CoA (RV = 9.4308, P = 0.0347, X) are core metabolites in lipid metabolism an essential metabolic process of CAD pathogenesis [42], which suggests that these newly discovered SGRCs are likely to be prime candidate biomarkers of CAD risks.

Table 1
The CAD pathogenic mechanisms of SGRCs in eight sub-networks.

[a]Functional categories	KEGG ID	[c]Cellular localization
Lipid and lipoprotein metabolism	C00116 C01885 C00641 C00422 C00165 C11136 C02112	C
Purine alkaloids	C00655	C
Oxidation–reduction	C03150 C00162 C00060	C
[b]Others	CE2884 CE6272 C11061 C00330 C00362	C
Lipid and lipoprotein metabolism	C00116 C00422 C00641 C01885 C00165	E
Immune and inflammation	C00060 C000162	E
Oxidation–reduction	C00162 C00060	E
Purine alkaloids	C00153	E
Thrombosis	C00035 C00325	G
Lipid and lipoprotein metabolism	G00063 G00081	G
Glucose metabolism	C00325 C01019 C00270	G
[b]Others	C04936 C04922	G
Lipid and lipoprotein metabolism	C02686 C00249 C00319 C12144	L
[b]Others	C02472 C06128 C00069 C06412	L
Renin–angiotensin system	C00020 C05993 C01344	M
Homocysteine metabolism	C00082 C00025 C05938 C00179 C01165 C03921 C01157 C03287 C00134	M
Immune and inflammation	C00134	M
Vascular smooth muscle cell abnormalities	C00410 C00025 C00535 C04295 CE0926 C03912	M
Lipid and lipoprotein metabolism	C00416 C00164 C00681	M
[b]Others	C00042 C00332 C02839 C00356 C00004 C00003 C00787 C01653 C00040 C09820 C00080 C00086 C01645	M
Endothelial integrity	C00021	N
Lipid and lipoprotein metabolism	C00399	N
[b]Others	C00080 C05544 C05545 C05546 C00019 C02415	N
Oxidation–reduction	C00390 C00399	N
Endothelial integrity	C14786 C14852 C06790 C11088 C13645 C14857 C14859 C14870 C00051	R
Lipid and lipoprotein metabolism	C05957 C04685	R
Renin–angiotensin system	CE5244	R
Oxidation–reduction	C14786 C14852 C06790 C11088 C13645 C14857 C14859 C14870 CE6235 C11304 C02320	R
[b]Others	C05953	R
Gender difference	C14786 C14852 C06790 C11088 C13645 C14857 C14859 C14870	R
Lipid and lipoprotein metabolism	C00040	X
Immune and inflammation	C09819	X
Oxidation–reduction	C00080 C00003 C00004 C09820	X

[a] Functional categories: lipid and lipoprotein metabolism, glucose metabolism, renin–angiotensin system, endothelial integrity, gender difference, vascular smooth muscle cell abnormalities, homocysteine metabolism, immune and inflammation, Purine alkaloids, thrombosis, oxidation–reduction and others.

[b] Others including generic compound participate in multiple metabolic reactions, nucleic acids, peptides and common amino acids.

[c] Cellular localization: C = cytoplasm, E = extracellular space, M = mitochondria, G = Golgi apparatus, R = endoplasmic reticulum, L = lysosome, X = peroxisome and N = nucleus.

2.4. Comparison between the RCM method and traditional approach

We carried out traditional approach to screen traditional genetic risk compounds (TGRCs) of CAD from the overall metabolic background. As a result, 89 TGRCs have been screened out (additional files Table S6), and 84(94%) of TGRCs were included in our SGRCs. Besides, 125 of SGRCs derived from RCM method were not identified by the traditional method (Fig. 2A). According to the distribution of TGRCs in organelles, most of TGRCs are predominantly in the mitochondria and endoplasmic reticulum, which is consistent with the results of the RCM method. We next compared the frequency of intra-organelle distributions of genetic risk compounds between the overlaps and differences. Interestingly, the overlaps were more likely to localize in single organelles (Fig. 3). However, the rest of SGRCs filtered out by traditional method were discovered to localize among multi-organelles (Fig. 3). These differences were largely due to the fact that although traditional method could extract metabolic reactions from the overall metabolic background, those reactions occurring in multi-organelles were then regarded as one whole entity during the identification process of genetic risk compounds. This might lead to the lack of identification of those compounds located in multi-organelles. On the contrary, the RCM method took sub-cellular localizations into account by separating EHMN into eight sub-networks and made the identification of potential organelle-specific SGRCs more reasonable and accurate.

In addition, literature retrievals were carried out to search for the evidence of disease relatedness of genetic risk compounds screened out by traditional and our methods. As a result, nearly one third of TGRCs (30) were found to be associated with CAD pathogenesis, all of which were also encountered in SGRCs. It should be noted that majority of all verified CAD-risk metabolites (73) found by our method were neglected by traditional method. To be more exact, the overlapping risk metabolites participated in several CAD-related functional categories including

Table 2
Frequency of cases and controls for each of two alleles at a SNP locus.

Allele	*X*	*Y*
Case	a	b
Control	c	d

[a] a–d represents the frequency of the specific allele X or Y, respectively.

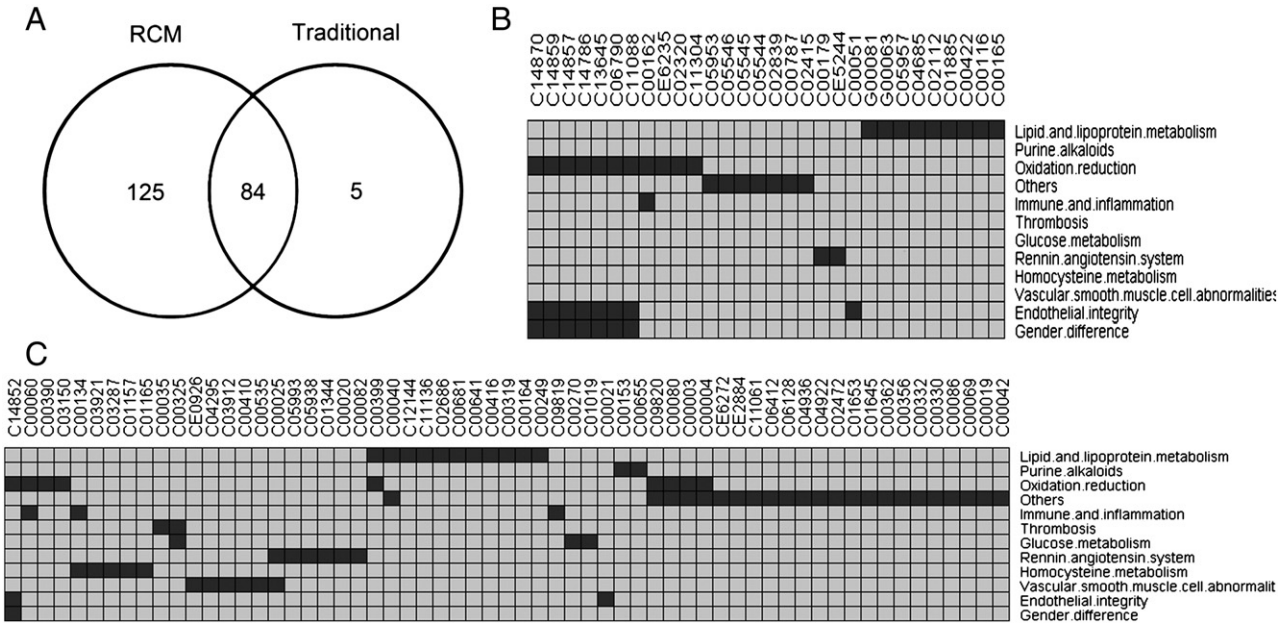

Fig. 2. Comparisons between the overlaps and differences of the two approaches. (A) Venn diagram for overlap metabolites identified by RCM and traditional method. (B) A heatmap of metabolits' functional distribution for literature-confirmed 30 metabolites both in RCM and traditional method. (C) A heatmap of metabolites' functional distribution for literature-confirmed 73 metabolites only identified by RCM. Black indicates high levels of functional enrichment, and gray indicates low levels.

lipid and lipoprotein metabolism, *oxidation–reduction*, *endothelial integrity and gender difference* (Fig. 2B). Noticeably, the literature confirmed that SGRCs identified only by the RCM method were also found to be related to other pathogenic processes of CAD (e.g., *renin–angiotensin system*, *vascular*, *smooth*, *muscle cell abnormalities and homocysteine metabolism*) (Fig. 2C), suggesting that the traditional method experienced lower power/performance in identifying genetic risk metabolites in the pathogenic processes of CAD.

3. Discussion

Coronary artery disease is considered to be caused by multiple genetic and environmental factors. In this paper, we proposed the RCM method which considered both the genetic factors and biologic networks to find CAD-related SGRCs. It should be noted that we did not extract all the reactions from metabolic networks as traditional methods, but introduced a more reasonable method by separating EHMN into eight sub-networks and dealt with those networks, respectively. Interestingly, we found 209 SGRCs distributed specifically among the 8 sub-networks, and most SGRCs concentrated on the mitochondria (19%), endoplasmic (29%) and peroxisome (23%). Unfortunately, current studies on sub-cellular localization of classes of metabolites (and enzymes) appeared to be incomprehensive. It still requires a lot of observations and experiments to accommodate metabolic localization resources. We will continue to focus on the development of this field in order to better decipher our results.

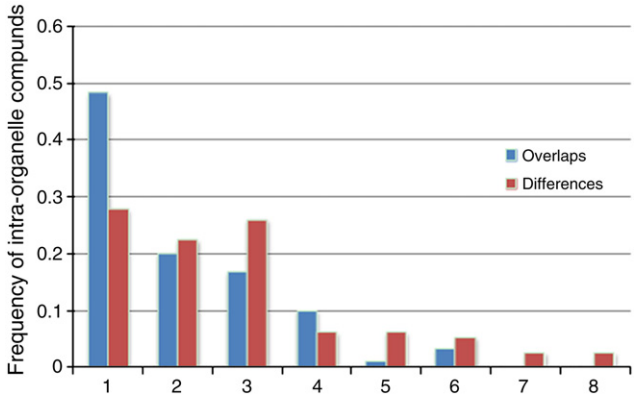

Fig. 3. A comparison of intra-organelle frequencies between overlaps and differences.

Through literature retrieval and functional classification, majority of SGRCs associated with coronary artery disease have been well verified. Furthermore, some SGRCs termed as conventional genetic risk factors are closely related to the pathogenesis of coronary artery disease. In addition, many SGRCs with higher heritabilities than conventional genetic risk factors suggested a strong correlation between genotype and phenotype. Comparing with traditional approach, the RCM method not only identified organelle-specific risk compounds more comprehensive and accurate, but also provided additional insights into the pathogenesis of CAD in the perspective of sub-cellular localization. These results have strongly illustrated that CAD-related SGRCs using the RCM method are reliable, which might provide more clues for traditional clinical diagnosis. However, some of the newly discovered SGRCs have not been confirmed yet, further experimental studies are necessary to assess the underlying mechanisms of these SGRCs in CAD.

As we know, CAD is likely to be caused by a series of factors, for example, genetic mutations, similar environment among family members, or the interaction between environmental factors and the traditional genetic factors [43]. At present, researches on the integration of genetic and metabolic networks are still limited. However, the results of our research not only found genetic risk related metabolites confirmed well in CAD, but also introduced a comprehensive list of functional categories of our SGRCs, which provided novel insights into CAD pathophysiology and genetics.

4. Materials and methods

4.1. Data source

The network was derived from the Edinburgh Human Metabolic Network, EHMN [14], which was reconstructed by integrating genome annotation information from different databases and metabolic reaction information from literature. EHMN contains nearly 3000 metabolic

reactions, which were reorganized into 70 human-specific pathways according to their functional relationships. In our study, the network was separated into 8 sub-networks (additional files Table S3) according to its localization information, including cytoplasm, extracellular space, mitochondria, Golgi apparatus, endoplasmic reticulum, lysosome, peroxisome and nucleus.

Genotype frequencies of tested SNPs of case–control samples were downloaded from the Welcome Trust Case Control Consortium (WTCCC) online system from the following samples using the 500K Affymetrix chip: 459,448 samples of CAD [15]. Location information of the human genes was acquired from the NCBI genome database (*Build 37.1*, Feb 2009). CAD related genes were downloaded from the CADgene database [16]. Human Metabolic pathway information was obtained from the KEGG database [17].

4.2. SNP significance analysis and risk evaluation

In this study, we used high-throughput SNP dataset as research source. Then we adopted a *t*-test and minimal allele frequency method to screen the frequency of cases and controls (Table 2).

By using the matrix table above, we calculated the statistics $\chi^2$and computed the corresponding probability p value of each SNP. Under the significance level of 0.05, a SNP set was preliminarily screened for the following study. The Pearson χ^2 formula was as follows:

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(b+d)(d+c)(c+a)} \tag{1}$$

An S_{risk} statistic can be acquired for each SNP if its significance level meets the threshold of 0.05 [18].

$$S_{\text{risk}} = \begin{cases} \log \dfrac{f_{\text{case}}(X)}{f_{\text{control}}(X)} & f_{\text{case}}(X) > f_{\text{control}}(X) > 0 \&\& P<0.05 \\ \log \dfrac{f_{\text{case}}(Y)}{f_{\text{control}}(Y)} & f_{\text{case}}(Y) > f_{\text{control}}(Y) > 0 \&\& P<0.05 \\ 0 & \text{others} \end{cases} \tag{2}$$

Here $f_{\text{control}}(X)$, $f_{\text{control}}(Y)$, $f_{\text{case}}(X)$,$f_{case}(Y)$ are the frequencies of cases and controls at the alleles *X* and *Y*, respectively. However, if the P value of a SNP was not significant, its risk value was set to be 0. For each SNP, theS_{risk} value was introduced to depict its relationship with CAD. The S_{risk}value could, in some sense, reflect the genetic effects of CAD.

4.3. Evaluation of gene risk

Most of functional elements (enhancers, repressors) are <500 kb away from genes; and most linkage disequilibrium blocks are less than 500 kb as well [19]. Therefore, for each gene termed as *g*, we choose the highest risk value among the SNPs which located within *g* as its genetic score. The formula was as follows:

$$risk(g) = \max_{1\le x\le m} \left\{ S_{\text{risk}_1}, S_{\text{risk}_2}, \ldots, S_{\text{risk}_x}, \ldots, S_{\text{risk}_m} \right\} \tag{3}$$

where *m* is the total number of SNPs mapped to *g*, S_{risk_x} is the risk value of the *x-th* SNP located in *g*.

4.4. Evaluation of reaction risk

Generally, the genotype of complex diseases is considered to be related with genetics and surrounding factors. Herewith, we calculated the average risk value of the gene g_i which was mapped into the reaction R_t.

$$R_t : A + B \xrightarrow{g_1, g_2, \ldots, g_i, \ldots g_n} C + D \tag{4}$$

For each R_t, we have defined its risk value as follows:

$$S_{R_t} = \frac{1}{n}\sum_{i=1}^{n} \{risk(g_i)\} \tag{5}$$

Here, *A*, *B*, *C* and *D* represent the products and substrates of a given reaction; $g_i (i = 1,2,\cdots,n)$ represents a regulation gene of a metabolic reaction; *n* is the total number of genes in R_t and $risk(g_i)$ is the risk value of the *i-th* gene in the reaction.

4.5. Evaluation of compound risk

Here we adopted an average risk effect to measure the genetic risks of metabolic compounds. For each compound, we evaluated the risk scoring of S_{compound} with relative value by calculating the average risk value of S_{reaction_k} which the compound participated in, and its formula was:

$$S_{\text{compound}} = \frac{\sum_{k=1}^{L} \left\{ S_{\text{reaction}_k} \right\}}{L} \tag{6}$$

In formula (6), L is the total number of reactions which the *compound* participated in.

4.6. Randomization test

In order to detect the significance of genetic risk compounds, we promoted 1000 times randomization to all sub-cellular networks, respectively. The randomization test was carried out for each compound; it was introduced to estimate the stability of the RCM method. This approach retained the structure of the metabolic sub-networks and randomized only the frequency of all the compounds. We adopted *Z*-score to screen the statistically significant risk compounds ($Z>3.14$; $P<0.05$).

$$Z = \frac{S_{\text{compound}_j} - \overline{\mu_{\text{compound}_j}}}{\sigma_{\text{compound}_j}} \tag{7}$$

In formula (7), S_{compound_j} represents the real risk value of the *j-th* compound, μ_{compound_j} is the mean risk value of the *j-th* compound for 1000 randomization and $\sigma_{\text{compound}_j}$ is the standard deviation of 1000 times randomization for each compound. The calculated p values from *Z*-scores were defined as measures of risk levels. In the next step, we screened SGRCs by the RCM method from 8 sub-networks with P values under a significant threshold of 0.05.

4.7. Algorithm of RCM method

The steps of the RCM method are shown as follows (Fig. 4):

I) Calculate risk values for all the SNPs whose P values meet the threshold of 0.05 according to the Pearson χ^2 statistic.
II) Screen and map SNPs to the corresponding genes in the EHMN.
  i) Map all the SNPs in step I to the corresponding gene that are located <500 kb away from *g*.
  ii) Select the maximum risk value S_{risk_x} as the genetic statistic value for $risk(g)$
III) Compute theS_{R_t} for the reaction R_t as formula (5)
IV) Compute the S_{compound} for the *compound* as formula (6)
V) Repeat steps III and IV to compute S_{compound} values for all the *compounds* in the network.
VI) Construct random networks for 1000 times, repeat steps III and IV to compute all the S_{compound} values in random networks.
VII) Compute *Z*-scores of all metabolites in the network as formula (7),

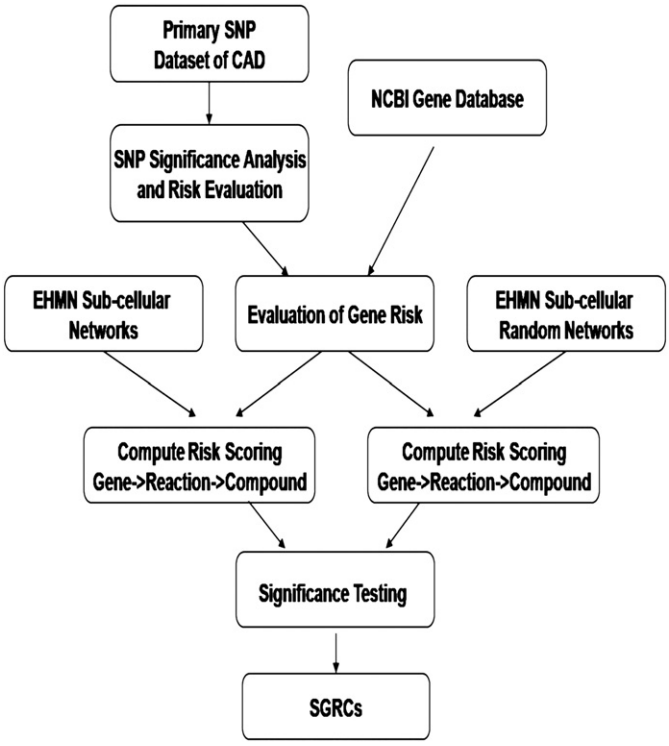

Fig. 4. Algorithm of the RCM method.

and the screened SGRCs with *Z*-scores above the threshold of 3.14 ($P<0.05$).

4.8. Functional enrichment analysis

Metabolite function enrichment in our study was calculated by the *MBRole* online system [20].

Acknowledgments

This work was supported in part by the Science & Technology Research Project of the Heilongjiang Ministry of Education (Grant No. 12511271) and the Master Innovation Funds of Heilongjiang Province (Grant No. YJSCX2011-338HLJ) and the National Natural Science Foundation of China (Grant NO. 30900837).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.07.013.

References

[1] R.S. Vasan, Biomarkers of cardiovascular disease, Circulation 113 (2006) 2335–2362.
[2] A. Bordbar, N.E. Lewis, J. Schellenberger, B.Ø. Palsson, N. Jamshidi, Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions, Mol. Syst. Biol. 6 (2010).
[3] K. Suhre, S.Y. Shin, A.K. Petersen, R.P. Mohney, D. Meredith, et al., Human metabolic individuality in biomedical and pharmaceutical research, Nature 477 (2011) 54–60.
[4] S.H. Shah, E.R. Hauser, J.R. Bain, M.J. Muehlbauer, C. Haynes, et al., High heritability of metabolomic profiles in families burdened with premature cardiovascular disease, Mol. Syst. Biol. 5 (2009).
[5] S.M. Purcell, N.R. Wray, J.L. Stone, P.M. Visscher, M.C. O'Donovan, et al., Common polygenic variation contributes to risk of schizophrenia and bipolar disorder, Nature 460 (2009) 748–752.
[6] P. Linsel-Nitschke, J. Heeren, Z. Aherrahrou, P. Bruse, C. Gieger, et al., Genetic variation at chromosome 1p13. 3 affects sortilin mRNA expression, cellular LDL-uptake and serum LDL levels which translates to the risk of coronary artery disease, Atherosclerosis 208 (2010) 183–189.
[7] D.L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M.E. Dolan, et al., Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS, PLoS Genet. 6 (2010) e1000888.
[8] B.S. Sutton, D.R. Crosslin, S.H. Shah, S.C. Nelson, A. Bassil, et al., Comprehensive genetic analysis of the platelet activating factor acetylhydrolase (PLA2G7) gene and cardiovascular disease in case–control and family datasets, Hum. Mol. Genet. 17 (2008) 1318.
[9] L. Chen, L. Zhang, Y. Zhao, L. Xu, Y. Shang, et al., Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways, Bioinformatics 25 (2009) 237.
[10] T. Hao, H.W. Ma, X.M. Zhao, I. Goryanin, Compartmentalization of the Edinburgh human metabolic network, BMC Bioinformatics 11 (2010) 393.
[11] B.D. Athey, J.D. Cavalcoli, H. Jagadish, G.S. Omenn, B. Mirel, et al., The NIH national center for integrative biomedical informatics (NCIBI), J. Am. Med. Inform. Assoc. 19 (2012) 166–170.
[12] A. Karnovsky, T. Weymouth, T. Hull, V.G. Tarcea, G. Scardoni, et al., Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data, Bioinformatics 28 (2012) 373–380.
[13] A. Zelezniak, T.H. Pers, S. Soares, M.E. Patti, K.R. Patil, Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes, PLoS Comput. Biol. 6 (2010) e1000729.
[14] H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, et al., The Edinburgh human metabolic network reconstruction and its functional analysis, Mol. Syst. Biol. 3 (2007).
[15] E. Zeggini, M. Weedon, C. Lindgren, T. Frayling, K. Elliott, et al., Wellcome Trust Case Control Consortium (WTCCC), McCarthy MI, Hattersley AT: replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes, Science 316 (2007) 1336–1341.
[16] H. Liu, W. Liu, Y. Liao, L. Cheng, Q. Liu, et al., CADgene: a comprehensive database for coronary artery disease genes, Nucleic Acids Res. 39 (2011) D991.
[17] M. Kanehisa, The KEGG database, in: Silico'Simulation of Biological Processes, 2002, pp. 91–103.
[18] M. Tian, M. Tang, H. Ng, P. Chan, Confidence intervals for the risk ratio under inverse sampling, Stat. Med. 27 (2008) 3301–3324.
[19] K. Wang, M. Li, M. Bucan, Pathway-based approaches for analysis of genomewide association studies, Am. J. Hum. Genet. 81 (2007) 1278–1283.
[20] M. Chagoyen, F. Pazos, MBRole: enrichment analysis of metabolomic data, Bioinformatics 27 (2011) 730.
[21] H. Ma, A.P. Zeng, Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, Bioinformatics 19 (2003) 270.
[22] H. Yamazaki, Y. Inui, C.H. Yun, F.P. Guengerich, T. Shimada, Cytochrome P450 2E1 and 2A6 enzymes as major catalysts for metabolic activation of N-nitrosodialkylamines and tobacco-related nitrosamines in human liver microsomes, Carcinogenesis 13 (1992) 1789.
[23] M. Corral-Debrinski, J. Shoffner, M. Lott, D. Wallace, Association of mitochondrial DNA damage with aging and coronary atherosclerotic heart disease, Mutat. Res. 275 (1992) 169–180.
[24] G.H. Werstuck, S.R. Lentz, S. Dayal, G.S. Hossain, S.K. Sood, et al., Homocysteine-induced endoplasmic reticulum stress causes dysregulation of the cholesterol and triglyceride biosynthetic pathways, J. Clin. Invest. 107 (2001) 1263–1292.
[25] A. Tenenbaum, M. Motro, E.Z. Fisman, E. Schwammenthal, Y. Adler, et al., Peroxisome proliferator-activated receptor ligand bezafibrate for prevention of type 2 diabetes mellitus in patients with coronary artery disease, Circulation 109 (2004) 2197–2202.
[26] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, et al., KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29.
[27] H. Watkins, M. Farrall, Genetic susceptibility to coronary artery disease: from promise to progress, Nat. Rev. Genet. 7 (2006) 163–173.
[28] N.J. Samani, J. Erdmann, A.S. Hall, C. Hengstenberg, M. Mangino, et al., Genomewide association analysis of coronary artery disease, N. Engl. J. Med. 357 (2007) 443–453.
[29] A.C. Lo Prete, C.H. Dina, C.H. Azevedo, C.G. Puk, N.H.M. Lopes, et al., In vitro simultaneous transfer of lipids to HDL in coronary artery disease and in statin treatment, Lipids 44 (2009) 917–924.
[30] A.S. Greenberg, R.A. Coleman, F.B. Kraemer, J.L. McManaman, M.S. Obin, et al., The role of lipid droplets in metabolic disease in rodents and humans, J. Clin. Invest. 121 (2011) 2102–2110.
[31] L. Villacorta, A. Azzi, J.M. Zingg, Regulatory role of vitamins E and C on extracellular matrix components of the vascular system, Mol. Aspects Med. 28 (2007) 507–537.
[32] S. Yamagishi, T. Imaizumi, Diabetic vascular complications: pathophysiology, biochemical basis and potential therapeutic strategy, Curr. Pharm. Des. 11 (2005) 2279–2299.
[33] G.K. Kolluru, S.C. Bir, C.G. Kevil, Endothelial dysfunction and diabetes: effects on angiogenesis, vascular remodeling, and wound healing, Int. J. Vasc. Med. 2012 (2012).
[34] M. Helfand, D.I. Buckley, M. Freeman, R. Fu, K. Rogers, et al., Emerging risk factors for coronary heart disease: a summary of systematic reviews conducted for the US Preventive Services Task Force, Ann. Intern. Med. 151 (2009) 496–507.
[35] D.M. Muoio, C.B. Newgard, Molecular and metabolic mechanisms of insulin resistance and β-cell failure in type 2 diabetes, Nat. Rev. Mol. Cell Biol. 9 (2008) 193–205.
[36] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, et al., PubChem: a public information system for analyzing bioactivities of small molecules, Nucleic Acids Res. 37 (2009) W623–W633.

[37] G. Assmann, Lipid Metabolism and Atherosclerosis, Schattauer, 1982.
[38] Y. Yamada, H. Matsuo, S. Warita, S. Watanabe, K. Kato, et al., Prediction of genetic risk for dyslipidemia, Genomics 90 (2007) 551–558.
[39] F. Haynes, in: Endothelial dysfunction by trichloroethylene-induced oxidative stress, MEHARRY MEDICAL COLLEGE, 2010.
[40] A. Hassan, B.J. Hunt, M. O'Sullivan, R. Bell, R. D'Souza, et al., Homocysteine is a risk factor for cerebral small vessel disease, acting via endothelial dysfunction, Brain 127 (2004) 212.
[41] J. Jaeken, H. Carchon, Congenital disorders of glycosylation: the rapidly growing tip of the iceberg, Curr. Opin. Neurol. 14 (2001) 811.
[42] T. Gordon, W.P. Castelli, M.C. Hjortland, W.B. Kannel, T.R. Dawber, High density lipoprotein as a protective factor against coronary heart disease: the Framingham study, Am. J. Med. 62 (1977) 707–714.
[43] F.H. Epstein, V. Fuster, L. Badimon, J.J. Badimon, J.H. Chesebro, The pathogenesis of coronary artery disease and the acute coronary syndromes, N. Engl. J. Med. 326 (1992) 242–250.